Boletín de Estadística e Investigación Operativa
Vol. 30, No. 1, Febrero 2014, pp. 66-77

Historia y Enseñanza

Jean-Baptiste Estoup and the origins of Zipf's law: a stenographer with a scientific mind (1868-1950)

Alain Lelu
Grupo KIWI (LORIA, Nancy)
Universidad de Franche-Comté
alain.lelu@univ-fcomte.fr

Abstract

Statistical distributions with a power law have been observed for over a century in many domains of social sciences, as well as in natural and life sciences. They are of utmost importance for those building models applicable to human activities (e.g. the "long tail" phenomena). We present here the life and accomplishments of J-B. Estoup, who was the first to notice this type of distribution in the language domain, and inspired the subsequent formulations by G.K. Zipf and B. Mandelbrot. This study, first presented at the seminar on the history of probabilities and statistics held at Ecole des Hautes Etudes en Sciences Sociales on December the 7th, 2007 in Paris, is also a family testimony, the author being the grandson of J-B. Estoup.

Keywords: History of statistics, Power-law distribution, Word frequencies Zipf"s law, Stenography
AMS Subject classifications: 01A60, 62-03, 62G30

1. A modo de introducción: un toque personal de historia familiar

Jean Baptiste Estoup es mi abuelo materno. Uno de sus nietos, Jacques Estoup, ha conservado la memoria familiar, profesional y científica de nuestro abuelo y me ha proporcionado muchos documentos y fotos, se lo agradezco muchísimo. Entre los descendientes de J.B. Estoup el interés por el lenguaje y los idiomas se halla muy difundido. Mi hermana Denise cursó la carrera de profesora de lengua y literatura española. Mi sobrino Bruno se apasionó por el estudio del chino y actualmente por el desciframiento de la escritura maya. (Bruno Delprat – SeDyL-CELIA, CNRS, Villejuif e INALCO, París – creó con Stepan Orevkov –

Institut de mathématiques, Université Paul Sabatier, Toulouse – el sistema mayaTEX, un sistema de composición tipográfica de textos jeroglíficos mayas para la computadora). Mi hermano Jean Paul, historiador, es experto en toponimia. Cuatro de mis tías y un tío fueron taquígrafos. Mi tío Henri Estoup puso a punto el primer prototipo de telex francés en los años 1939, comercializado en 1948 por SAGEM. En cuanto a mí: dediqué toda mi trayectoria universitaria al análisis de datos textuales.

Figura 1: Foto de Jean-Baptiste Estoup en 1940

Tengo muy pocos recuerdos de mi abuelo: vuelvo a verlo tendido en su cama, enfermo, con su gran bigote blanco. Yo tenía entonces 5 años.

Mi niñez fue poblada de estos signos misteriosos que mi madre utilizaba para escribir la lista de sus compras o sus recetas de cocina, de palabras oscuras como Metagrafía directa Prévost-Delaunay, Duployé integral, y de lejanos ecos de riñas incomprensibles de la familia. Sabía que mi abuelo había sido un gran taquígrafo, nada más.

Los años pasan hasta el día de 1994 cuando, al leer el libro *Statistique textuelle* (Lebart y Salem, 1994), descubro que lo citan así como a Benoît Mandelbrot. Consideran a J.B. Estoup como uno de los padres fundadores de los estudios de las frecuencias de las palabras en el lenguaje.

> «...Sucedió que el primero – que yo sepa – en ocuparse de las frecuencias relativas de las palabras en el discurso fue un taquígrafo de ingenio científico, Jean-Baptiste Estoup. Sus resultados han sido incomparablemente ampliados por Georges Kingsley Zipf, que en Harvard University enseñó una rara mezcla de locas elucubraciones y de hechos muy importantes, despreciados por sus coetáneos por ser demasiado difíciles de clasificar...» (Mandelbrot, 1968).

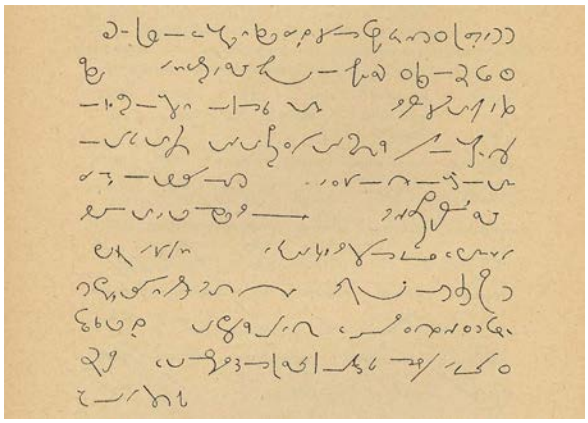

Figura 2: Ejemplo de texto taquigrafiado

Las charlas con mi primo Jacques Estoup, los documentos que me trajo así como las investigaciones genealógicas de mi hermano Jean Paul me han permitido completar el conocimiento de nuestro abuelo.

En la primera parte de la ponencia voy a evocar la vida y la obra de Jean-Baptiste Estoup. Luego, hablaré más precisamente de la ley de Zipf y de algunos de los innumerables trabajos que siguieron sus pasos.

1.1. La juventud de J.B. Estoup: ¡ cap de piteu! (testarudo en el habla *commingeois*)

J.B. Estoup nació en 1868, en una familia establecida desde hacía mucho tiempo en el Comminges, en el valle de Luchon; eran comerciantes, artesanos, maestros de escuela, corredores de libros (era un oficio tradicional de la región), mesoneros en Gaud y luego en Luchon durante la moda del termalismo en la primera mitad del siglo XIX, con buena o mala suerte.

Hace estudios clásicos en el Seminario de Gourdan-Polignan, cerca de Saint-Gaudens. Se gradúa de Bachiller en letras en Toulouse y luego en Pau se alista en el ejército francés, en la artillería. Ahí descubre la taquigrafía y la estudia con pasión con un compañero suboficial ¡pagará a uno de sus hombres para dictarle discursos parlamentarios! Aunque no le daban sino unas pocas licencias para tomar parte en concursos de taquigrafía y a pesar de fracasar varias veces, persevera (una foto lo muestra en Luchon taquigrafiando el discurso de inauguración de la estatua del valle del Lys).

Después de darse de baja del ejército en 1896, se instala en Paris. Ahí aprueba el concurso de taquígrafo de la Cámara de diputados. En 1899 le toca taquigrafiar el segundo proceso contra Dreyfus, para el diario Le Figaro, con René Havette, historiador de la taquigrafía, amigo suyo aunque usaban métodos taquigráficos diferentes: R. Havette practica el Prévost-Delaunay y J.B. Estoup el Duployé simplificado (la hija de R. Havette, Andrée, se casará con Jean-Henri el hijo de

J.B. Estoup). Participa activamente en los concursos y congresos del Instituto Taquigráfico de Francia y es nombrado secretario general de la Unión de las Sociedades Taquigráficas de Francia.

Se casa en 1897 con su prima Henriette a pesar de la oposición de su padre. En pocos años nacen los primeros de sus siete hijos. Según la costumbre burguesa de aquel entonces, los ponen al cuidado de una nodriza de los altos valles de Luchon, en Benqué-Dessus. Mi madre tenía mucho cariño por Memé de Benqué, y a los tres años no sabía ni una palabra de francés, se expresaba en el habla del Comminges.

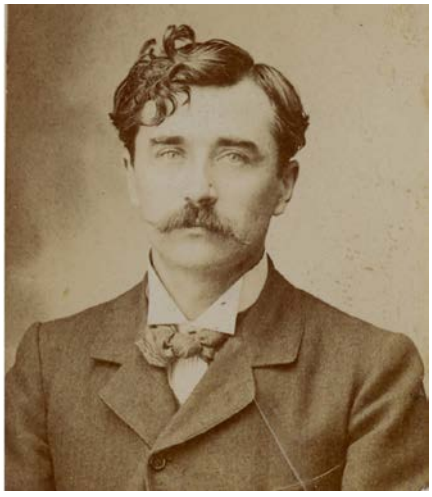

Figura 3: Foto de Jean-Baptiste Estoup en los años 1890

1.2. La madurez: un progresista (¡y feminista!) en «la Belle époque»

J.B. Estoup comunicó su pasión por la taquigrafía a su esposa (por la que tenía una viva admiración), a su hijo y a cuatro de sus hijas. Obró con mucha perseverancia para abrirles las puertas del oficio de taquígrafo de discurso, reservado en aquella época a los hombres. Su hija Marguerite, pianista virtuosa y campeona de taquigrafía a los 22 años con 190 palabras por minuto, no logró entrar en el Senado (¡no podía entregar una libreta militar!) en 1924 la Corte Internacional de Justicia de La Haye reconoció sus méritos contratándola, y luego la ONU en Nueva York hasta su jubilación. Mi abuela Henriette por su lado militó contra toda discriminación, y si bien nunca logró entrar en la Cámara de Diputados fue taquígrafa de dos Consejos regionales y de la Sociedad de las Naciones.

La pareja trabó contactos internacionales, en particular durante el congreso internacional de taquigrafía organizado con motivo de la Exposición universal de 1900. Tuvieron como buenos amigos al taquigrafista alemán R. Fuchs y al turco S. Hudaverdoglu.

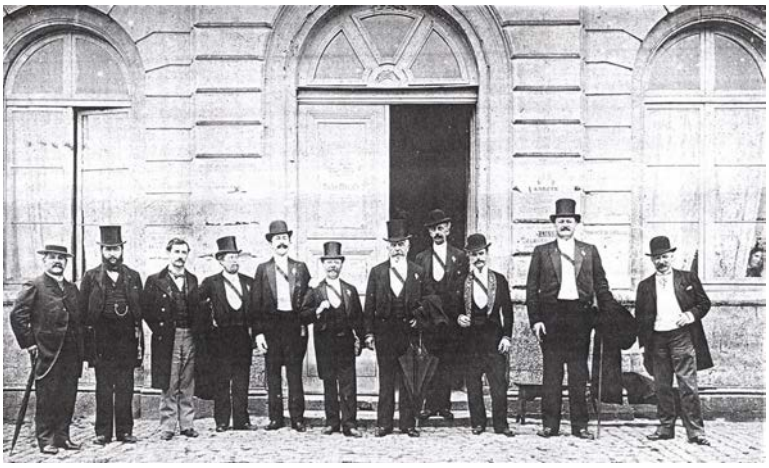

Figura 4: El taquígrafo (tercero), el conserje y los diputados de los Pirineos.

2. Un espíritu racional e inventivo

2.1. De Duployé a Estoup

El sacerdote Emile Duployé concibió en los años 1860 una escritura fonética para instruir a los analfabetos, una escritura popular, simplificada y sin el afán de rapidez de escritura. Esta generosa utopía no tuvo el éxito esperado, pero fue adoptada por taquígrafos que pronto le añadieron métodos de aceleración (metagrafía). Se publicó en 1895 un Curso Parlamentario. A partir de 1897 J.B. Estoup trae su contribución: una progresión más lógica y una mejor coordinación entre las diferentes partes de la obra enriquecen la edición de 1898.

Pero este curso no le satisface, critica un sinnúmero de ardides heteróclitos propios de los que ejercen el oficio. Entonces concibe los principios que debían llevar hacia una coherencia del todo y que con terquedad no dejará de promover durante toda su vida

- Descartar lo que no corresponde a *reglas racionales*. Enseñar directamente la metagrafía sin pasar por la integral - algunas reglas, en vez de un sinnúmero de ardides. Primacía a los datos sacados de la experiencia: Frecuencia de los sonidos, de los enlaces, de las palabras; medición del número de los alzamientos de la pluma y cambios de dirección por minuto. Por fin inventa el concepto de gama y crea gamas taquigráficas (de 50 palabras por minuto a 140); su corpus, con varios temas -política, economía, comercio, ciencias y técnicas, derecho...- alcanzará las 112000 palabras.

2.2. Un análisis científico

Se deduce de sus recuentos y de diversas experiencias que

- Una palabra francesa taquigrafiada consta de un promedio de 3,5 cambios de dirección del trazo.

- El límite fisiológico es de 800 cambios de dirección por minuto.

- De ahí un límite práctico de 230 palabras por minuto, muy superior a las 120 hasta 170 palabras por minuto de los oradores «normales». Por esto J.B. Estoup se opone a una abreviatura exagerada y propone abreviar las palabras frecuentes haciendo más legibles las palabras poco frecuentes. En efecto la «traducción» es EL problema de los taquígrafos, porque la escritura fonética es ambigua, y la metagrafía lo es más aún. El taquígrafo que traduce debe tener el contexto con precisión en la mente, pero también una buena cultura general y dotes literarios para traducir a una lengua escrita correcta.

2.3. Un frenesí de recuentos

Los cálculos de N-gramas de signos son viejos como la criptografía. Un frenesí de recuentos de fonemas se apodera de los taquígrafos en la segunda mitad del siglo XIX, más particularmente de

- La Société Française de Sténographie (1896): examen de 33000 palabras.

- Friedrich W. Kaeding y unos veinte colegas de la Universidad de Dresde examinan en 1898 un corpus de 11 millones de palabras alemanas. Desdichadamente, este trabajo se hará inutilizable en la óptica ulterior de Zipf ...

Aparecen recuentos de *palabras*: Reverent J. Knowles (en Londres, 1904): 100000 palabras; R.C. Eldridge (en Nueva York, 1911): 35000 palabras.

3. Dos estudios de recuento de las palabras del francés

J.B. Estoup empieza en aquel entonces (antes de 1912) dos estudios sobre las palabras del francés. Su corpus es el de sus gamas taquigráficas ya calibradas por grupos de unas 1000 palabras (una o dos páginas).

3.1. Estudio de incremento lexicográfico

Se trata de contestar a la pregunta: ¿Cuántas palabras diferentes están comprendidas en el discurso? En un corpus de 30000 ocurrencias de palabras

- 20000 son entregadas a Auguste Touzeau, profesor de taquigrafía

- 10000 (¿en un principio 14000?) son coordinadas por J.B. Estoup.

Después de examinar los datos, observa en el gráfico aquí arriba una relación (poco más o menos) hiperbólica entre el número de palabras nuevas y diferentes por una parte, y el efectivo acumulado de las palabras por otra parte. De ahí

extrapola un límite práctico de unas 3000 palabras diferentes en un corpus de 60000 ocurrencias de palabras, vocabulario restringido del orador medio que el taquígrafo de discurso encuentra en realidad frente a él. El aprendiz de taquígrafo deberá aprenderlas y entrenarse hasta trazarlas automáticamente.

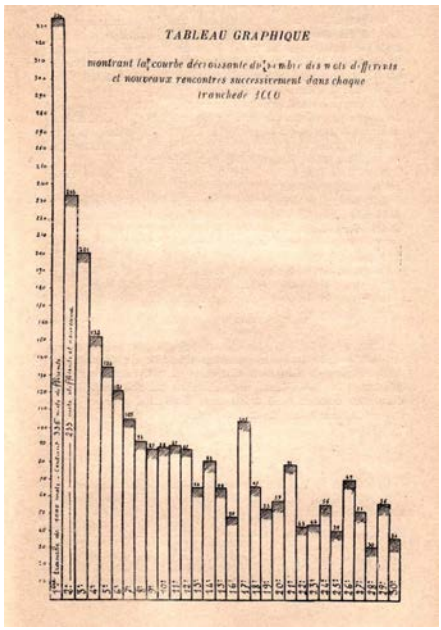

Figura 5: Decremento, por grupos de 1000 ocurrencias sucesivas, del número de palabras diferentes (primer grupo), luego nuevas y diferentes.

3.2. Estudio de las repeticiones

Se trata de contestar a la pregunta: ¿Cuántas veces se repiten las palabras más frecuentes?, para guiar, informar objetivamente a los que toman la iniciativa de crear o mejorar los signos.

El estudio se funda en un corpus de 20000 palabras:

- J.B. Estoup saca una lista de listas de palabras repetidas n veces, por orden decreciente de repeticiones; agrupa ciertas palabras homófonas corrientes *(et, est, ai)*, artículos *(le, la, les)*, las formas conjugadas de los verbos, las formas del singular y del plural de los nombres, las formas masculinas y femeninas de los adjetivos.

- Más acá de 7 repeticiones, las palabras ya no son detalladas.

- Ningún gráfico (pero más tarde, Zipf lo traducirá directamente de modo gráfico: número cumulado de palabras repetidas × número de repeticiones, con coordenadas log-log).

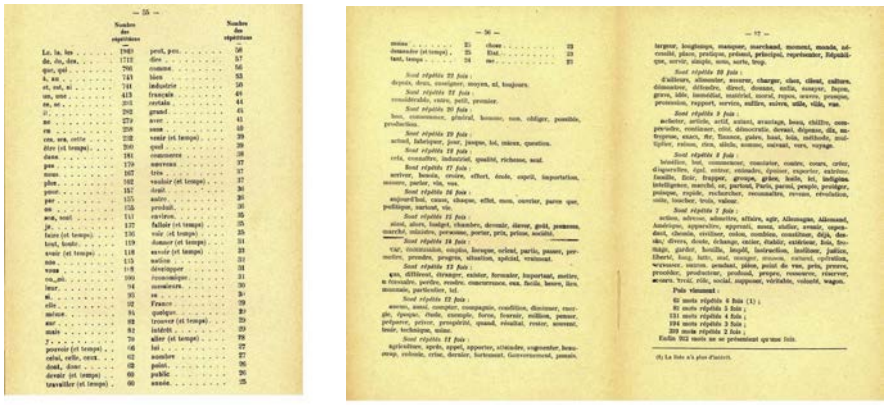

Figura 6: Lista de las palabras más frecuentes.

Estos dos estudios se publicaron en 1916, en el folleto teórico (Estoup, 1916) que acompañaba la cuarta edición de sus dos libros de gamas taquigráficas. No tenemos más que la séptima edición, en la familia. La cuarta parece que se conserva, con su folleto de documentos adjuntos, en la biblioteca municipal de Nueva York, donde B. Mandelbrot ha podido consultarla. La tercera en 1912 contiene una alusión al límite de las 3000 palabras corrientes, de modo que podríamos considerar que los recuentos son anteriores a 1912 (Petruszewycz, 1973). Subsiste gran parte de los escrutinios parciales por grupos. Es un trabajo repartido entre varias personas, como lo atestiguan la diversidad de las escrituras y algunos nombres: André Fauconnier, Max Muller ...

J.B. Estoup, de formación literaria clásica, nunca se ha interesado por la teoría de por sí; pero ha aplicado con ahinco al perfeccionamiento de la técnica (¿o del arte?) un proceso racional de observación, de ahí su iniciativa de recuentos y de experimentación (si era necesario probaba sus innovaciones en sus alumnos y sus niños...). Su acción ha tendido constantemente hacia la depuración progresiva de un sistema enturbiado en un principio por mucha arbitrariedad, y también hacia la enseñanza con un aprendizaje progresivo, pero de un solo grado: las mismas reglas se aplican a la progresión en el aprendizaje de las palabras más difíciles (por ejemplo, la transcripción de los diptongos) y en la velocidad que es necesario alcanzar. Este método le ha proporcionado el éxito – sus gamas taquigráficas se publicaron a unos cien mil ejemplares – pero también le ha granjeado un creciente alejamiento de sus colegas más conservadores.

4. Estoup y Zipf

George Kingsley Zipf, nacido en 1902 en los Estados Unidos y muerto en 1950, empezó una tesis doctoral de filología en Alemania de 1924 a 1929, en Bonn, y luego en Berlín; la defendió en Harvard en 1929. En su carta a J.B. Estoup (18 de junio de 1927), le pide con suma cortesía que le mande una copia de los cuadros

del artículo del taquígrafo J.B. Illio en la revista *L'Eclair sténographique* de 1911 a propósito de una lista de pólizas de frecuencia, « polices de fréquence » (en francés en la carta), «Lauthafigkeit» en alemán (insiste en los fonemas). Hay poca esperanza de averiguar si se trataba más bien de frecuencias de palabras ya que en la BNF (Bibliothque Nationale de France) esta revista se halla clasificada «HU», fuera de uso. Lo verosímil es que se trata de fonemas, objeto de su tesis.

Es liegt mir sehr auch die Ergebnisse dieser Bearbeitung zu haben, und deshalb belästige ich Sie mit meiner Bitte.

Haben die "Eclair" für das Jahr 1911, und würden Sie so sehr freundlich sein mir eine Abschrift der darinenthaltenen Tabellen zu schicken? Und wenn der Aufsatz nicht allzulang ist, könnten Sie nicht Ihren Herrn Sekretär bitten, auch die wichtigsten Züge (resumé) des Aufsatzes mir abzuschreiben und zuzuschicken. Ich würde seine Bemühungen gern bezahlen, und zwar was Sie und er für richtig halten.

Ich kann Ihnen nicht sagen, Herr Doctor, wie sehr Sie mich dadurch verpflichten würden, und wie unendlich dankbar ich Ihnen wäre. Und falls Sie irgend etwas hier in Berlin zu erledigen haben, würde ich es als eine Ehre auffassen, wenn Sie mich damit auftragen würden.

Verzeihen Sie mir, bitte, wenn ich Ihnen Deutsch anstatt Französisch schreibe. Obgleich ich gut Französisch kann, möchte ich mich nicht vor einem hervorragenden Sprachkenner wie Sie blamieren.

Mit verbindlichem Gruss in vorzüglichster Hochachtung, bleibe ich,

Ihr ergebener

George Zipf

Membre de la Société Linguistique de Paris

Herrn Doctor J. B. Estoup
Paris

Figura 7: Página 2 de la carta de G.K. Zipf a J.B. Estoup, del 18 de junio de 1927.

4.1. Primera formulación de la ley de Zipf

- 1929: tesis de filología comparada, *Relative Frequency as a Determinant of Phonetic Change*, en que J.B. Estoup forma parte de las personas a las que agradece.

- 1932: la obra: *Selected studies on the principle of relative frequency in language*, Zipf (1932)

. Cita a J.B. Estoup.

. Traduce directamente en gráfico log-log la expresión « lista de listas », del estudio Estoup de las frecuencias de las palabras:
* Abscisa (j): efectivo de las clases de frecuencias de palabras (número de palabras repetidas 1 vez, 2 veces,... , por ejemplo : las 4 palabras *considérable, entre, petit, premier* son repetidas 21 veces.)
* Ordenada (b): frecuencias de estas palabras

. ... aplicándola a otros datos: latín de Plauto, inglés (recuento Eldridge), palabras e ideogramas chinos.

. Constata la relación «universal» para el 95 % de las palabras: $jb^2 = constante$, fórmula «exactamente idéntica a la de la gravitación»...

. ... pero hace una trampa con las palabras de efectivo 1, que deberían tener una frecuencia fraccionaria para obedecer a su ley: ¡no las representa!

4.2. Segunda formulación de la ley de Zipf

- 1935: *The psycho-biology of language*, Zipf (1935).
- Nueva formulación que utiliza los rangos de las palabras clasificadas por efectivos decrecientes:
  - Abscisa (i): rangos de las clases de frecuencias de palabras (la palabra más frecuente, la segunda más frecuente,...).
  - Ordenada (b): frecuencias de estas palabras.
- Constata la nueva relación «universal» **frecuencia** $\times$ **rango = constante**
- ... que incluye esta vez las palabras de efectivo 1.

Sabemos ahora que estas dos formulaciones son equivalentes, Haitun (1982):

- La primera puede ser expresada como una ley de densidad de probabilidad: $P(j) \sim j^b$ o dicho de otro modo, expresa la probabilidad que una palabra sea presente j veces en el corpus.
- La segunda como una ley de densidad de probabilidad que una palabra tenga el rango i por orden de frecuencias decrecientes: $F(i) \sim i^a$
- Con, como puente entre los dos, la relación: $b = 1 + 1/a$

4.3. Después de Estoup y de Zipf

Las observaciones y las formalizaciones progresivas de Estoup y de Zipf se insertan en una corriente general de descubrimientos de estas «leyes de potencia» en numerosos sectores de las ciencias del hombre y de la vida (economía, redes sociales, genómica ...), así como en las ciencias de la materia (longitud de los ríos, meteorología ...), desde el final del siglo XIX hasta hoy (Barbut, 2003). Han suscitado muchos intentos de modelización explicativa. Podemos citar en el campo de la lengua:

- Benoît Mandelbrot en 1960 via la entropía de Shannon (Mandelbrot,1960).
- Harald Baayen en 2001 via distribuciones LNRE, Large Number of Rare Events (Baayen, 2001).

5. Epílogo: un apasionado en una época de pasiones

Para J.B. Estoup los años de la guerra de 1914-1918 fueron penosos: sesiones de noche interminables en la Cámara de diputados, dificultades para vivir en París durante la guerra, fallecimiento a los diez años de su hija Françoise y obligación de renunciar a sus amistades internacionales, como la de R. Fuchs en Alemania. Sin embargo crea en 1917 el boletín *La Vérité Sténographique*, sostén de sus convicciones y de su enseñanza, que va a sobrevivirle hasta 1992. Sigue más que nunca sus actividades múltiples de investigador independiente (publicación en 1916 de su estudio de las frecuencias), profesor, editor y prosélito de su « métagraphie directe duployenne ».

En el año 1918 empieza la separación de los partidarios del método de Duployé, entre los que se negaban a reanudar el contacto con los taquígrafos «enemigos», y los que lo deseaban ardientemente, como J.B. Estoup y sus amigos. En 1924 éstos abandonan el Institut Sténographique de France y fundan el Institut International de Métagraphie Duployé. J.B. Estoup se jubila en 1929 pero sigue con ardor sus actividades.

Después de la segunda guerra mundial y con las primeras preocupaciones por el porvenir de la taquigrafía, era urgente que las diferentes corrientes de la taquigrafía Duployé se unieran para perdurar en la enseñanza pública francesa: nació el gran proyecto de codificación, terminado en enero de 1949 por Henri Estoup, hijo de Jean-Baptiste. A los talentos de su padre que unía la práctica a la teoría, añadió el de conciliador. Jean-Baptiste Estoup, que no favorecía los compromisos, murió en abril de 1950.

Agradecimientos A Jacques y Geneviève Estoup, Jean-Paul Lelu, Bruno Delprat, y Denise Delprat por su trabajo de traducción, sin olvidar lo que debemos a los estudios de la Señora Micheline Petruszewycz (1973).

Referencias

[1] Baayen R.H. (2001). *Word Frequency Distributions*, Kluwer Academic Publishers, Dordrecht (Paises bajos).

[2] Barbut M. (2003). Homme moyen ou homme extrême ? De Vilfredo Pareto (1895) Paul Lévy (1935) en passant par Maurice Fréchet et quelques autres. *Journal de la Société Franaise de Statistique*, **144**, n° 1-2.

[3] Estoup J.B. (1916). *Gammes sténographiques : méthode et exercices pour l'acquisition de la vitesse*, 4e édition rev. et aug., 151p., 20 rue Gassendi, París (Francia).

[4] Haitun S.D. (1982). Stationary scientometric distributions. Part II. Non-Gaussian nature of scientific activities. *Scientometrics*, **4** N°2, 89-104.

[5] Lebart L., y Salem A. (1994). *Statistique Textuelle*, Dunod, París (Francia) (Francia).

[6] Mandelbrot B. (1960). On the theory of word frequencies and on related markovian models of discourse. En: *Structure of Language and its Mathematical Aspects*, 190-219. Ed. Roman Jakobson (Symposia in Applied Mathematics XII, Providence), R.I.: American Mathematical Society, New York (EE.UU.)

[7] Mandelbrot B. (1968). *Les constantes chiffrées du discours, Le Langage*, Encyclopédie de la Pléiade, vol XXV, Gallimard, París (Francia).

[8] Petruszewycz M. (1973). L'histoire de la loi d'Estoup-Zipf : documents. *Mathématiques et Sciences Humaines*, **44**, 41-56.

[9] Zipf G.K. (1932). *Selected Studies of the Principle of Relative Frequency in Language*, Cambridge (Mass., EE.UU.).

[10] Zipf G.K.(1935). *The Psycho-Biology of Language*, Cambridge (Mass., EE.UU.).

Acerca del autor

Alain Lelu es profesor en Ciencias de la Información y sus líneas de investigación son el análisis de datos textuales y la factorización matricial.